DOCUMENT RESUME

ED 104 241 HE 006 363

AUTHOR Reid, John C.
TITLE On the Reliability and Validity of Interviewers of Medical School Applicants.
PUB DATE [75]
NOTE 15p.

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Admission (School); Admission Criteria; *Higher Education; *Interviews; *Medical Schools; Reliability; Research Projects; *Validity

ABSTRACT

This study investigates how closely different interviewers agree when ranking the same applicants, and determines the "correctness" of their rating of applicants. Interview data for applicants to a medical school for two recent consecutive years were examined. For the first year, 573 applicants were interviewed; for the second year, 675 were interviewed. Twenty-six physicians were interviewed in the first year, and 38 interviewed in the second year, with 20 interviewed both years. In the first year, 146 pairs of interviewers interviewed the same candidates. In year 2, 73 of the 238 pairs were retained. Results indicated that most interviewers were both reliable and "correct" or valid. Interviewers who were identified as candidates for unreliability were inspected to see if these interviewers gave ranking incongruents with the admissions committee, thus providing ready identification of those interviewers who are reliable, and whose judgments are in accord with validity criteria. (MJM)

ED104241

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

On the Reliability and Validity of Interviewers of Medical School Applicants

John C. Reid

School of Medicine

University of Missouri - Columbia

Introduction

The selection of medical students from a large pool of applicants is an extraordinarily difficult process because of the large number of highly qualified applicants. Undergraduate grade point average (GPA) is the best single predictor of success in medical school (1), and GPA along with scores on the Medical College Admissions Test (MCAT) are seriously considered by admissions committees. Even when applicants with lower GPA's and MCAT scores are rejected, the remaining number of seemingly qualified applicants is larger than the number of available positions in medical school class. To help resolve this dilemma, most admission committees personally interview each member in this smaller, selected pool. At the medical school in the present study, at least two physicians, members of the admissions committee, interview each applicant. It would be hoped first of all that the two interviewers closely agree on their rating of each applicant, and second that their rating is "correct."

If the two interviewers agree in their rating of the applicants, then the two interviewers may be considered to be reliable. If the in-

AE006363

interviewers' judgments are not only reliable but also "correct" or "true," then their judgment is also valid. Considering the expense of both candidates' and interviewers' time and the career decisions that are being made, the need for determining the reliability and validity of interviewers' judgments is obvious. Furthermore, the need for such a study is strengthened by the fact that despite the admitted importance of the concepts of reliability and validity, few reports of these coefficients for admissions committees' interviews have appeared in the literature (2).

The purpose of the present study was to investigate how closely different interviewers agree when ranking the same applicants, and also to determine the "correctness" of their rating of applicants.

In particular, this study determined (1) the degree to which pairs of interviewers assign the same applicant the same rank, (2) whether each particular interviewer ranked his interviewees similar to the rank assigned by the admissions committee working in toto (the first validity criterion), and (3) whether each particular interviewer's ranks correlated with later student success in medical school (the second validity criterion).

Procedure

Interview data for applicants to a medical school for two recent consecutive years (called year one and year two) were examined. For

year one, 573 applicants were interviewed; for year two, 675 were interviewed. Twenty-six physicians interviewed in year one, and 38 interviewed in year two, and 20 of these interviewed both years. Physicians were non-randomly assigned to interview applicants. Interviewers could ask the applicant any questions they felt were useful, but interviewers specifically needed to gather data to answer questions on a structured interview form. For year one, the structured form contained 17 questions having three to five responses. For year two, the structured form had 10 five-response questions. Typical questions on the form required the interviewer to rank the applicant on motivation, or work load outside of studies. After the interview, and upon considering the applicant's GPA, MCAT, and letters of recommendation, the interviewer assigned the applicant an overall evaluation rank of 5, 4, 3, 2, or 1 for year one, or 5, 4, 3, or 1 for year two, with 5 being the highest. Interviewers were not aware of other interviewer's ratings.

In year one, 146 pairs of interviewers interviewed the same candidates. Sixty-three of these pairs interviewed more than three applicants and were retained in the reliability analysis. In year two 73 of 238 pairs were retained. The same interviewer typically interviewed applicants in common with five or six other interviewers. All interviewers rating more than three applicants were retained in the validity analysis; out of 26 satisfied this criterion in year one and 30 out of 38 satisfied this in year two. Although a statistic based

on a small number of observations may not justify a hard decision, it may call attention to a need for more information. The median number of applicants ranked by these interviewers was 37.5 and 38 for the two years.

It is worthwhile to review the method of determination of reliability and validity coefficients before proceeding to the results of the study. Had all interviewers interviewed all applicants, then such multivariate methods of reliability as proposed by Cronbach et al., (3) would be applicable. However, since interviewers only interviewed a few applicants each, and that non-randomly, the applicant by interviewer data matrix is somewhat similar to those matrices discussed by Shoemaker (4), except that procedures for estimation of such a variance-covariance matrix have not been worked out beyond some important preliminary work by Timm (5) and Chan (6).

To determine the agreement between each pair of interviewers ranking the same applicants, four distance formulas were computed: a Euclidian distance D^2 (7), its square root D, a distance which was the sum of absolute values of deviations between interviewers, and a distance which was the sum of a (0,1) loss function. The binary loss function was defined as 0 unless the rank differences between interviewers exceeded unity. For each pair of interviewers, all four distance functions were divided by the number of applicants rated by the pair.

It is important to realize that statistics such as correlation and

anova destroy some of the information that the distance functions retain and therefore would be less valuable as reliability coefficients in this study than measures of distance to measure profile similarity of interviewers. Two examples may suffice to illustrate this point. Two interviewers may evaluate applicants Tom, Dick and Harry as follows: Tom: 5, 3; Dick: 3, 3; Harry: 1, 3. An anova would produce no significant differences between the mean ratings of the two interviewers, yet the interviewers clearly rate the three applicants differently. A second example: Tom: 5, 3; Dick: 4, 2; Harry: 3, 1. A correlation would equal unity but would destroy the important mean differences between interviewers.

Although the distance functions do not have a limited range, as does the correlation coefficient, the retention of raw score units is an interpretative advantage, rather than a disadvantage.

Cronbach and Gleser (7) and Rulon et al. (8) have discussed the similarity between D^2, Mahalanobis D. and the discriminant function.

For a pair of interviewers to have a high inter-distance value, that is, to disagree on their ratings of the same people, one or both interviewers could have made errors in judgment. It could be the fate of a "correct" interviewer to be paired with an "erroneous" interviewer. Therefore, each member of an interviewer pair having a mean distance function value, D, of >.71 was regarded as a candidate for unreliability, since such a D value indicated that these interviewers would on

the avarage evaluate a candidate more than half a category apart.

To determine whether interviewers' evaluations were "correct" (valid) using the first criterion, rank order correlations rather than distances were calculated between the interviewers' rankings and the committee rankings, since the two rankings were on different instruments. (A slight error occurs with this scheme because each interviewer is a member of the admissions committee. The error is similar to that in an item--total score correlation.)

Interviewers having rank-order correlations (rho) of <.6 with the admissions committee final rating were regarded as not valid; interviewers having rank-order correlations of $\geq$.6 were regarded as valid.

Interviewers can be thought of as being in one cell of a 2 × 2 table, the columns of which are labeled candidates for unreliability (yes, no), and the rows of which are labeled satisfactory validity (yes, no). Decisions about what to recommend for interviewers in each of the four cells will now be discussed.

Interviewers who were classified as not being a candidate for unreliability and who also had high validity coefficients should be retained on the admissions committee.

Interviewers who were classified as not being a candidate for unreliability but who had low validity coefficients should probably have their performance reviewed. However, if applicants who these interviewers would have turned down were accepted into medical school and

had difficulty, then the admissions committee is erroneously ignoring the insights of these interviewers.

Interviewers who were classified as being an candidate for unreliability and who also had satisfactory validity coefficients can probably be retained on the admissions committee. A useful statistic to help in making this decision derives from the fact that each interviewer typically interviewed applicants in common with five or six other interviewers, and a distance for each pair can be computed. This statistic is the ratio of the number of times an interviewer was paired with an interviewer having disparate ratings (symbolized by U for unlike) to the total number of interviewers (T) he was paired with. If the U/T ratio was 1, then the interviewer disagreed with every other interviewer he was paired with. If the U/T ratio was 0, then the interviewer agreed with every other interviewer he was paired with, and would not be a candidate for unreliability. It is possible, of course, for a set of interviewers to agree with each other, yet they all be erroneous ("incorrect" or invalid). If the U/T ratio is >.5, then the interviewer may not be sufficiently stable in his judgments to warrant retention on the admissions committee without further training.

Finally, interviewers who were both candidates for unreliability and also had low validity coefficients probably should be dropped from the admissions committee, particularly if their recommendations are not substantiated by later student performance in medical school.

The criterion of the decision of the total admissions committee is valuable because it can be computed using all applicants to medical school, not just accepted students. As time goes on, though, a second criterion of performance in medical school becomes available, which necessarily derives from a smaller group. For each interviewer for each year, rank order correlations (rho) were computed between interviewer's rating, the mean number of times a student got honors in a course, mean delayed grades, mean subjective rating as a house officer, and for year one students, the NBME part I total score. Interviewers having all four correlations positive for accepted students were judged valid on the second criterion; interviewers having one or more of the four correlations negative were judged invalid on the second criterion. For those interviewers whose ratings had been judged as invalid using the first criterion of admissions committee decision, the progress of the particular students they rated was examined to see if that interviewer's initial rating was substantiated by that students' progress in medical school.

Results

Rank-order correlations between the least squares, absolute value, and loss distance formulas were obtained. Correlations between least squares and absolute value distances were .99 and .91 for years one and two, between least squares and loss were .80 and .81, and between abso-

lute value and loss were .75 and .60. Differences in decisions based on the use of differing distance functions will not be further discussed here.

For year one, of all candidates for unreliability, 40% had interviewer-committee correlations <.6 (were not valid), and 60% had interviewer-committee correlations of ≥.6 (were valid). Of the 60% who had valid, but possibly unreliable ratings, only one had a U/T ratio of ≥.5; most had U/T ratios of .2 or .1. A (U/T) ratio of .5 means that that interviewer disagreed with half of the other interviewers who rated the same applicants. One interviewer had a validity coefficient of <.6, but he had not been identified as a candidate for unreliability because he had fewer than four interviews in common with any other interviewer.

For year two, of all candidates for unreliability, half had interviewer-committee correlations <.6 (were not valid), and half had correlations ≥.6. Of this latter half, only 2 had U/T ratios ≥.5. Of the interviewers who had unsatisfactory validity coefficients but who were not candidates for unreliability, half had fewer than four interviews in common with any other interviewer, and thus would not have been identified as a candidate for unreliability.

The design of a longitudinal study permits the comparison of earlier performance with later performance. If a judge's performance is constant across time, then increasingly greater confidence is obtained that the judge is being correctly classified. Interviewers who remain

reliable and valid across years should be retained on the admissions committee; those who remain unreliable and not valid across years might be given another assignment. In the present study, two interviewers remained in the non-valid category for both year one and year two.

It should be mentioned that certain students were re-interviewed when the committee could not reach agreement. These re-interviews were not included in the present analysis.

Data from those interviewers identified as invalid by the first criterion of admissions committee decision were examined to see if students whom they rated low did poorly in medical school, and if students whom they rated high did well. Two examples will illustrate possible outcomes. Table 1 (a) was produced by Dr. X, 1 (b) by Dr. Y.

Students were classified on whether they got a mean course rating of 4 or higher on a 7-point scale. A student doing this well is rarely, if ever, in trouble in that course.

If a student was rejected by interviewer X, the probability was .57 that he would do satisfactory work in medical school. If a student was accepted by interviewer X, the probability was .67 that he would do satisfactory work in medical school.

Thus, although interviewer X was classified by the first criterion as an invalid interviewer, he may, on balance, be marginally acceptable.

If a student was rejected by interviewer Y, the probability was one that he would do satisfactory work in medical school. If a student was

accepted by interviewer Y, the probability was .33 that he would do satisfactory work in medical school. Thus, although these data for Y are based on only 8 students, the data support the original classification of Y as an invalid interviewer. Similar analyses were done on honors, delayed or failing grades, and NBME - 1 scores, although they are not reported here.

Summary and conclusions

In this two-year study, the evaluation rankings given by interviewing physicians to 1391 medical school applicants were investigated for similarity of rankings between interviewers (reliability), for similarity of judgments between interviewers and the admissions committee (validity), and for similarity of interviewers' judgments and later student performance.

Most interviewers were both reliable and "correct" or valid as operationally defined herein. Ratings by interviewers who were identified as candidates for unreliability were also inspected to see if these interviewers gave rankings incongruent with the admissions committee.

Three reasons could account for the rankings of those interviewers classed as candidates for unreliability who also have low validity coefficients. The first reason could be any kind of error such as interviewer, instrumental, recording, or interviewer-interviewee interaction. Some of this error could be decreased by the review of interviewing

principles. The second could be that these interviewers correctly perceived some positive or negative trait of the interviewee that others failed to see or failed to be persuaded of. The third could be that interviewers are not rating applicants on traits relevant to medical school performance.

Attrition occurred in computing both reliability and validity indices because some interviewers rated only a few applicants. This attrition could be reduced in future studies if interviewers were required to interview at least 20 applicants.

The method described permits ready identification of those interviewers who are reliable, and whose judgments are in accord with validity criteria. It is most important for an institution to be aware of the reliability and validity of one of the major portions of the admissions process and to rectify any correctable components once they have been identified.

Table I

Data from interviewers classified as invalid on the criterion of admissions committee decision

(a)

Interviewer X's rating	Number of Students Rated	Number of Students Getting >4 in Mean Course Ratings
Acceptable	12	8
Unacceptable	7	4
	19	12

(b)

Interviewer Y's rating		
Acceptable	6	2
Unacceptable	2	2
	8	4

References

1. Brading, P. L. The Relationship between Success in Medical School and both Selected Academic and Non-academic Prediction Factors. Unpublished Ph.D. dissertation, University of Southern California, 1971.

2. Rimm, A. A., Pazdral, P., and Sine, J. Methodological Study of Rating Applicant Interviews. J. Med. Educ., 43:1085, 1968.

3. Cronbach, L. J., et al. The Dependability of Behavioral Measurements. New York: Wiley, 1972.

4. Shoemaker, D. M. Principles and Procedures of Multiple Matrix Sampling. Inglewood, California: Southwest Regional Laboratory for Educational Research and Development, August 1971.

5. Timm, N. H. The Estimation of Variance-covariance and Correlation Matrices from Incomplete Data. Psychometrika, 35:417-437, 1970.

6. Chan, L. S., and Dunn, O. J. The Treatment of Missing Values in Discriminant Analysis--I. The Sampling Experiment. J. Amer. Stat. Assn., 67:473-477, 1972.

7. Cronbach, L. J., and Gleser, G. C. Assessing Similarity Between Profiles. Psychological Bulletin, 50:456-473, 1953.

8. Rulon, P. J., et al. Multivariate Statistics for Personnel Classification. New York: Wiley, 1967.